Semantic and Style based Multiple Reference Learning for Artistic Image Aesthetic Assessment

Tengfei Shi[a,c,d], Chenglizhao Chen[b,*], Xuan Li[d] and Aimin Hao[a]

[a]Beihang University, Beijing, 100083, China
[b]China University of Petroleum (East China), Qingdao, 266555, China
[c]Qingdao Research Institute, Qingdao, 266104, China
[d]Peng Cheng Laboratory, Shenzhen, 518055, China

ARTICLE INFO

Keywords:
artistic image aesthetic assessment
multiple reference learning
graph convolutional network
deep learning

ABSTRACT

Artistic Image Aesthetic Assessment (AIAA) is an emerging paradigm that predicts the aesthetic score as the popular aesthetic taste for an artistic image. Previous AIAA takes a single image as input to predict the aesthetic score of the image. However, most existing AIAA methods fail dramatically to predict the artistic images with a large variance of artistic subjective voting with only a single image. People are good at employing multiple similar references for making relative comparisons. Motivated by the practice that people considers similar semantics and specific artistic style to keep the consistency of the voting result, we present a novel Semantic and Style based Multiple Reference learning (SSMR) to mimic this natural process. Our novelty is mainly two-fold: (a) Similar Reference Index Generation (SRIG) module that considers artistic attribution of semantics and style to generate the index of reference images; (b) Multiple Reference Graph Reasoning (MRGR) module that employs graph convolutional network (GCN) to initialize and reason by adjusting the weight of edges with intrinsic relationships among multiple images. Our evaluation with the benchmark BAID, VAPS and TAD66K datasets demonstrates that the proposed SSMR outperforms state-of-the-art AIAA methods.

1. Introduction

With the breakthrough in image style transfer Gatys, Ecker, and Bethge (2016); Huang and Belongie (2017); Park and Lee (2019) and the rapid development of Artificial Intelligence Generated Content (AIGC) Zhang, Lei, Jia, Ma, and Hu (2018), users can now effortlessly create artistic images with varying styles and content. The online sharing and creative design of generative artistic images have gained increasing popularity, even being utilized in artistic competitions, where they have achieved first place Gault (2022). However, controlling generative detail remains challenging, and assessing artistic image aesthetics has become an essential requirement. Artistic Image Aesthetic Assessment (AIAA) mimics human aesthetic perception to predict the aesthetic quality of artistic image, which has potential applications in enhancing the effectiveness of virtual face design Seow, Lim, Phan, and Liu (2022); Zeng, Zhao, Zhang, Liu, and Li (2022); Patotskaya, Hoyet, Olivier, Pettré, and Zibrek (2023), scene generation Yi, Huang, Tripathi, Hering, Thies, and Black (2023), sketch completion Voynov, Aberman, and Cohen-Or (2023); Cheng, Chen, Chiu, Tseng, and Lee (2023), etc. In recent years, driven by the rapid progress of AIGC, the evaluation of aesthetic quality in artistic images has emerged as a prominent research area.

A great number of Image Aesthetic Assessment (IAA) models Li, Zhi, Shi, Yang, Xu, Li, and Guo (2023); Zhang, Luo, Zhang, Wu, Wang, and Shen (2023); Zeng, Cao, Zhang, and Bovik (2019); Talebi and Milanfar (2018) have reported, yet the methodology for AIAA remains relatively unexplored. The primary objective of AIAA is to develop models that effectively capture generic features for mimicking people's aesthetic perception when evaluating artistic visuals. Current generic AIAA approaches Yi, Tian, Gu, Lai, and Rosin (2023) are primarily employed to address two types of aesthetic tasks: aesthetic binary classification and aesthetic score regression. The aesthetic binary classification task intends to categorize images into two classes: high aesthetics and low aesthetics. In the aesthetic score regression task, the model predicts numerical scores denoting the average aesthetic perception vote for a given image. These two tasks are interrelated, as aesthetic binary classification can be derived from the aesthetic scores using defined thresholds.

Early AIAA models extracted hand-crafted features to mimic human aesthetic perception, such as color histograms Nishiyama, Okabe, Sato, and Sato (2011), statistical methods Zujovic, Gandy, Friedman, Pardo, and Pappas (2009) and pyramid based methods Parker (2010).Although hand-crafted aesthetic features include explicit photographic composition and physical meanings, they lack a comprehensive representation of the abstract and diverse attributes of artistic image aesthetics. This limitation arises from the highly subjective image aesthetics, making it challenging to define abstract concepts through hand-crafted features precisely. With the development of deep neural networks (DNNs), the representational capabilities of DNNs have significantly improved AIAA models. Existing AIAA models heavily depend on large annotated datasets and an end-to-end learning manner. Furthermore, these methods have proposed to improve the learning and representational abilities of the network, such as extracting multi-patch features, integrating the deeply learned features with hand-crafted features Zhang, Miao, Zhang, and Yu (2020), and

*Corresponding author
ORCID(s): 0000-0002-4260-3905 (T. Shi)

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

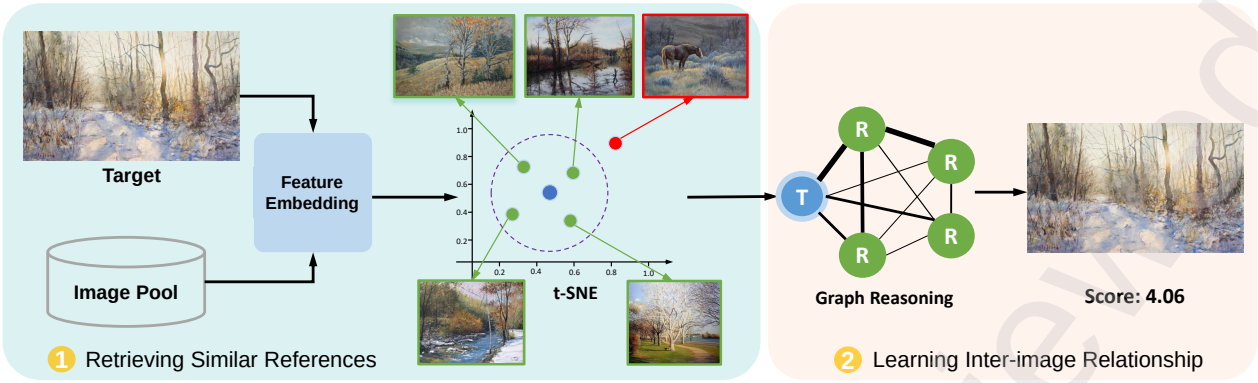

Figure 1: Illustration of the key compenets for our approach. To capture fine-grained artistic features for aesthetic represention, we first retrieve multiple reference images with similar features of aesthetic attributes. The first 4 nearest samples with target in t-SNE embedding space are then selected to input the latent space for efficient graph reasoning to predict the aesthetic score.

exploring the relationship between specific style features and general aesthetic features Yi et al. (2023). These efforts aim to strengthen the model's capabilities by investigating local relationships or integrating multiple features to enhance aesthetic representation. However, relying solely on individual images fails to precisely represent artistic aesthetic attributes such as semantics and style, thus being unable to alleviate the influence of subjectivity effectively.

People often make more consistent aesthetic judgments based on similar images with relative comparison. Specifically, these similar images include similar artistic attributes, e.g., scene semantic information and artistic painting styles, etc. These related works have recently found application across a range of aesthetic-related tasks. He et al. He, Zhang, Xie, Jiang, and Ming (2022) introduced the concept of extracting scene features from aesthetic images as priors, integrating them into IAA for assessing aesthetic quality. Yi et al. Yi et al. (2023) proposed to combine artistic style and generic aesthetic features from images to evaluate artworks. These methods show that it is crucial for AIAA to incorporate semantic and stylistic features to enrich its representation. However, current AIAA methods rely directly on feature extraction and fusion, which may not provide a fully explored on employing artistic attributes like human visual perception. Kan et al. Kan, Cen, Li, Vladimir, and He (2022) explores a graph neural network to characterize the local correlation structure of embedding features for outperforming image retrieval results, demonstrating the effectiveness of GNN. Therefore, we propose to leverage multiple references and graph reasoning to enhance artistic features and make reasonable aesthetic decisions. In Fig. 1, we illustrate the key process in two stages: retrieving similar references and learning inter-image relationships. The first stage is retrieving the target with an image pool to find informative references. We visualize the target and references in t-SNE embedding space, which measures the distance of semantic and style features to retrieve the reference index. In Fig. 1, the blue node is the target and four green nodes have similar semantic "forest" and style "oil painting", but the red node has a different semnatic "horse" and painting style of "watercolor". In the learning inter-image relationship stage, targets and references are used as graph nodes to build graph structure. We facilitate relative comparisons through graph reasoning to adjust the weights of edges instead of relying solely on feature fusion. Furthermore, graph reasoning can adjust the weight with different nodes incompletely depending on the similarity of the artistic aesthetic attributes. This approach demonstrates that informative reference images and employing graph reasoning to build rich representations greatly enhance the AIAA model and improve the results of aesthetic regression.

Inspired by these insights, we propose a multiple reference learning approach called Semantic-Style Multi-Reference (SSMR), as shown in Fig. 2, which consists of two components: the Semantic-Reference Image Retrieval (SRIG) module and the Multi-Reference Graph Reasoning (MRGR) module. Firstly, we introduce a novel reference image retrieval strategy, SRIG, which retrieves more similar artistic images as references. The SRIG module considers semantic and style attribute feature distances of artistic images to construct the reference index. By calculating the feature distances between the current observation target image and the images in the training set (image pool) and normalizing the accumulated distances, the SRIG module obtains the nearest samples as references. Second, we propose Multi-Reference Graph Reasoning (MRGR) to ensure the effectiveness of multi-sample learning. By inputting multiple reference images, MRGR initializes the edge weight values in the graph structure based on feature distances. Graph convolutional network reasoning infers the intrinsic relationship among multiple samples. As the graph convolutional network learns, MRGR enables the model to adjust appropriate relation weights with the similarity of the semantics and style between the reference images and the target image.

The main contributions of our work are three-fold:

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

- We present a novel Semantic and Style based Multiple Reference learning (SSMR) to explicitly relate the aesthetic perception to similar semantics and style attribution.
- We propose a Similar Reference Index Generation (SRIG) to construct semantic and style related reference index and Multiple Reference Graph Reasoning (MRGR) to reason the weight of intrinsic relationships among multiple artistic images.
- We evaluate the state-of-the-art IAA and AIAA approaches on the BAID, VAPS and TAD66K dataset. Our model performs favorably against on all the metrics, which demonstrates the validity of our model.

2. Related work

2.1. Image Aesthetic Assessment

Image aesthetic assessment (IAA) attempts to automatically evaluate the aesthetic level for binary classification, aesthetic score prediction or aesthetic score distribution prediction. IAA adapts to the highly subjective nature of different individuals who may have distinct aesthetic preferences. Most existing IAA methods rely on selecting local patch and fusion global view Lu, Lin, Jin, Yang, and Wang (2014); Lu, Lin, Shen, Mech, and Wang (2015); Ma, Liu, and Wen Chen (2017); Sheng, Dong, Ma, Mei, Huang, and Hu (2018) to improve the performance of predicting the aesthetic score. RAPID Lu et al. (2014) incorporated heterogeneous inputs, including a global and local view to predict the classifier of aesthetic level. DMA-Net Lu et al. (2015) further updated multiple patches and aggregate strategies to improve the performance of image aesthetic categorization. A-Lamp Ma et al. (2017) improved adaptive selected multi-patches and layout-aware attribute graphs for image aesthetic assessment. MPada Sheng et al. (2018) proposed an attention-based mechanism to adjust the weights of each patch during the training process to improve learning efficiency.

Latest IAA studies consider Graph Convolutional Network (GCN) to construct the relationship of multiple patches Liu, Puri, Kamath, and Bhattacharya (2020); She, Lai, Yi, and Xu (2021). RGNet Liu et al. (2020) built a region composition graph of multiple patches and reasoned relationships to boost the performance of aesthetics assessment. However, RGNet does not show performance gains in predicting aesthetic scores and score distribution. She et al. She et al. (2021) present a Hierarchical Layout-Aware Graph Convolutional Network (HLA-GCN) to capture layout information. However, HLA-GCN employed a Fully Convolutional Network (FCN) to extract spatial features and build the graph structure which increases computational costs.

2.2. Artistic Image Aesthetics Assessment

There are relatively few AIAA methods, where earlier traditional methods Amirshahi and Denzler (2017); Guo, Kurita, Asano, and Asano (2013) extract handcrafted features and train Support Vector Machine (SVM) for classification. There are relatively few AIAA methods, where earlier traditional methods Amirshahi and Denzler (2017); Guo et al. (2013) extract handcrafted features and train Support Vector Machine (SVM) for classification. Recently Zhang et al. Zhang et al. (2020) employed a deep multi-view parallel convolutional neural network (DMVCNN) to learn aesthetic features for Chinese ink paintings. Recently Zhang et al. Zhang et al. (2020) employed a deep multi-view parallel convolutional neural network (DMVCNN) to learn aesthetic features for Chinese ink paintings. In general, AIAA methods have not been adequately studied. Different from the above works, we argue that different artistic attributions need to retrieve different semantic and style related reference features, and construct both style-specific and aesthetic relationships to evaluate artworks.

3. Method

In this section, we first provide an overview of the proposed Semantic-Style Multi-Reference learning (SSMR). Then, we give the problem formulation and loss function of the artistic image aesthetic assessment task. Finally, we present two key modules: Semantic-Reference Image Retrieval (SRIG) for generating reference index with semantic and stylistic similarities; and Multi-Reference Graph Reasoning (MRGR) for learning the relationships of aesthetic attribution among multiple features.

3.1. Overview

Motivated by overcoming the inherent limitations of a single sample, the proposed SSMR aims to model multiple samples learning by performing graph reasoning on multiple reference images. The architecture of the proposed framework is shown in Fig. 2. Given the artistic image as the "target", we conduct the SRIG module to generate a reference index with the similarity of semantic and artistic style. Considering both semantic and style features extracted by artistic attribution embedding (Sec. 3.4), the SRIG module extracts the target and all training datasets for similarity metric (Eq. 3) to measure the multiple nearest samples as reference. The proposed framework initializes edge weights of the graph structure and performs graph reasoning on all samples to construct their intrinsic relationships via the multiple reference graph learning (MRGL) module. By performing graph reasoning among multiple samples, the proposed framework can model semantic and style of different visual elements on multiple images. Furthermore, we propose a multiple samples reference learning scheme to aggregate significant artistic attribution in a latent space, where a set of reference images with similar semantics and style can be projected onto a single feature for representation. Finally, target and reference features are fused as the aesthetic representation to predict the aesthetic score of the artistic image.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

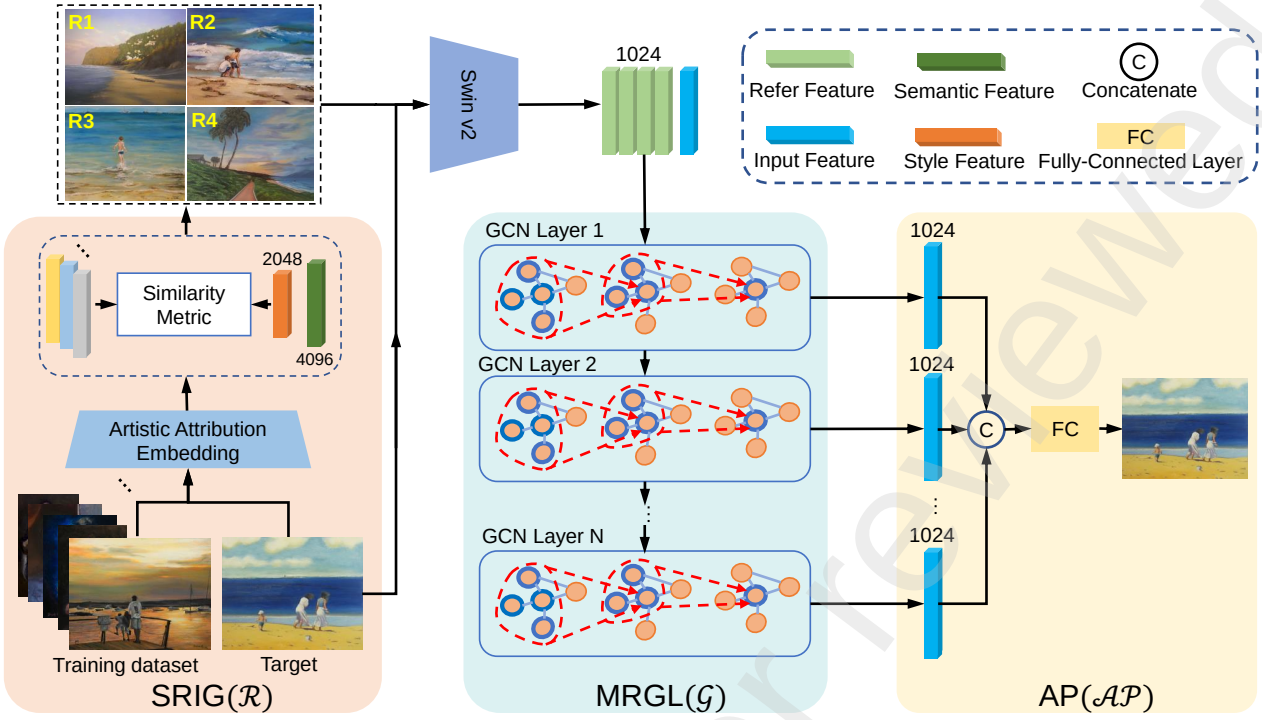

Figure 2: An overview of the proposed methodology: Similar Reference Index Generation (SRIG), Multiple Reference Graph Learning (MRGL) and Aesthetic Predict (AP). SRIG aims to metric the similarity of input feature and training dataset and in turn selects the suitable samples for generating reference index. MRGL reasons multiple samples in a graph structure with different node and weighted edge gradually adapts the semantic and style of artwork during the assessment. AP makes final aesthetic decision based on fine-grained aesthetic features. Sec. 3 provides further details.

3.2. Artistic Image Aesthetic Assessment Preliminaries

For the IAA problem, previous approaches have primarily focused on three tasks: binary classification, score regression, and score distribution with abundant labels of aesthetic score distribution for supervised learning. However, the AIAA dataset only provides aesthetic scores for artistic images used for score regression. Based on the previous AIAA methods, our objective is to employ deep learning networks $\mathcal{M}$ for accurately predicting aesthetic scores $\hat{y}$ of artistic images x. Compared to IAA methods that predict aesthetic distributions, the AIAA task is more challenging due to less supervised information and higher subjectivity.

Let us consider that $D_{art\_aesthetics} = \{x_i, y_i\}$ is an artistic aesthetic dataset, where x_i is the input artistic image and y_i denotes the average score of artistic aesthetic assessment. We aim to train a new model $\mathcal{M}$, based on the reference image R, consisting of a small set of artistic aesthetic dataset $D_{art\_aesthetics}$. For an artistic image x_i, the reference image dataset R_i is comprised by the m images, which is selected with the most informative semantic feature f_{sem} and artistic style feature f_{style}, serving as the optimal reference set. For each artistic image x_i and corresponding reference set R_i, it is easy to obtain the semantic feature f_{sem} and style feature f_{style} by the backbone network. Our model $\mathcal{M}$ consists of three modules: reference selection $\mathcal{R}$, graph reasoning $\mathcal{G}$ and aesthetic prediction $\mathcal{AP}$. First, the module $\mathcal{R}$ selects the most informative reference set r_i from $D_{art\_aesthetics}$, which are denoted by $R_i = \mathcal{R}(f_{sem}, f_{style}|D_{art\_aesthetics})$. Then, the module $\mathcal{G}$ employs the reference set R_i to construct the graph structure $\hat{f} = \mathcal{G}(\mathcal{N}, A, W)$, node $\mathcal{N} \in \mathbb{R}^{m+1}$ is the input image x_i and its m references R_i, A is initialized edges. By reasoning and updating the weight W of the relationship with multiple images as reference learning, the enhanced feature $\hat{f}$ can be obtained. Finally, the module $\mathcal{AP}$ makes use of artistic image and reference enhanced features $\hat{f}$ to predict the score of aesthetic quality, which is expressed as $\hat{y} = \mathcal{AP}(\hat{f})$. The $\hat{y}$ is the predicted aesthetic score, and we can adopt reference image set R to optimize the model $\mathcal{M}$:

$$\min_{\phi(x,y)\in D_{art\_aesthetic}} \mathcal{L}(\mathcal{M}(x|R;\phi), y), \tag{1}$$

where ϕ means the model parameters, $\mathcal{L}$ is the loss function.

Fig. 2 illustrates the overall framework, including three components: similar reference index generation (SRIG), multiple reference graph reasoning (MRGR), and aesthetic prediction (AP). Considering samples x_i, each aesthetic score with a value ranging from 0 to 10, we normalize the scores for stable training. The normalized scores are obtained by:

$$y_{norm\_i} = \frac{y_i - min(y)}{max(y) - min(y)}, \tag{2}$$

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

where $min(y)$ and $max(y)$ mean the maximum and minimum value of the aesthetic score in the training dataset. During the inference phase, our framework directly predicts the normalized scores, denoted as y_{norm}. The predicted result can be obtained by applying Eq. 2 to get the actual scores.

3.3. Key Insights of the Proposed Approach

It is different from assessing abstract and subjective artistic images by CNNs directly. Naively predicting the incoming test samples Talebi, Amid, Milanfar, and Warmuth (2020) could be problematic for both approaches: the single input is different to (a) extract instance-wise feature that is actually implicit or abstract contents to predict y, i.e., the "contents" is complex for artistic image to obtain by the model; and (b) recognise specific styles by only a single sample, which neglects the intrinsic relation of different artistic styles Graham, Hughes, Leder, and Rockmore (2012).

When assessing artistic image aesthetics, people tend to make relative comparisons among similar scenes instead of an absolute evaluation. Thus, the key insights of our approach are to find suitable references and make relative comparisons. In contrast to natural images, artistic images include more abstract and diversified artistic styles to represent their aesthetics. Therefore, identifying suitable references needs to be considered for the inherent abstract and artistic style within artistic images. Thus, we select reference images based on two attributes: similarity of visual semantics and correlation of artistic styles.

After obtaining reference images, we employ a graph convolutional network to facilitate relative comparisons and extract enhanced features. Traditional approaches often depended on additional features or score discrepancies for relative comparisons, which resulted in limited interpretability. In contrast, our method enables a direct observation of changes in connection weights among different nodes during graph construction. As illustrated in Fig. 2, nodes that exhibit higher similarities in both features and scores acquire increased weights, thereby strengthening feature representation. Conversely, nodes displaying lower feature similarities and notable score differences experience reduced connection weights, resulting in a more limited impact on feature improvement. This mechanism equips our model with abundant reference data, enabling efficient aesthetic evaluations through constructing enhanced features. As a result, it establishes the foundation for aesthetic reasoning.

3.4. Similar Reference Index Generation

As mentioned in Sec. 3.3, recent AIAA algorithms have only considered individual samples to learn specific styles and aesthetic features of images. However, we believe that this common approach may not successfully capture complete style representation F_{style} and semantic representation F_{sem} of the artistic image x_i in the training samples.

In principle, for a given artistic image x_i, accurate extraction of its semantic feature F_{sem} and style feature F_{style} through the model can enable the prediction of the label y_i. However, in the highly subjective artistic image aesthetic dataset $D(x, y)$, relying solely on a single image to extract accurate style and aesthetic features is insufficient. Direct feature extraction may fail to adequately correlate related information for extracting the semantic feature F_{sem} and style feature F_{style} from sample x_i.

To mitigate the challenges posed by the subjective and abstract in artistic images, we propose to construct a multiple reference index from the training set to correct the impact of subjectivity and abstractness. Instead of extracting style feature F_{style} and aesthetic feature $F_{aesthetic}$ from a single image directly, we first introduce Similar Reference Index Generation (SRIG), which considers both semantic and style features to build reference index R with similar artistic attribution:

$$\mathrm{Sim}_{aes}(\mathbf{I}_t, \mathbf{I}_i) = \left\| \left(\frac{\hat{\mathbf{F}}_t}{\|\hat{\mathbf{F}}_t\|_1}, \frac{\hat{\mathbf{F}}_i}{\|\hat{\mathbf{F}}_i\|_1} \right) \right\|_2, \tag{3}$$

where $\|\cdots\|_1$ denotes the l_1-norm, $\hat{\mathbf{F}}_t$ and $\hat{\mathbf{F}}_i$ are the feature of target $\mathbf{I}_t$ and $\mathbf{I}_i$ in training dataset, $\|\cdots\|_2$ means euclidean distance, ImageNet pre-trained ResNet50 and VGG19 are used to extract image semantic and style features Simonyan and Zisserman (2014) includes F_{sem} and F_{style} respectively.

To obtain a more reasonable metric for selecting the reference index, we introduce a weight parameter λ to adjust the influence of semantic and style features on the retrieval process. Specifically, we propose the following approach for reference image retrieval:

$$D(\mathbf{I}_t, \mathbf{I}_i) = \lambda * \mathrm{Sim}_{sem}(\mathbf{I}_t, \mathbf{I}_i) + (1-\lambda) * \mathrm{Sim}_{style}(\mathbf{I}_t, \mathbf{I}_i), \tag{4}$$

where $D(\mathbf{I}_t, \mathbf{I}_i)$ means the i-th image reference index based on fusion feature distance. λ is the weight affecting the style feature distance. By adjusting λ, we choose the appropriate metric weight to obtain a suitable reference index.

By computing the feature distances between the current observation and all samples in the training set, normalizing and aggregating the distances, we obtain the reference index $R(m)$ for the m closest samples in terms of semantic and style features from the current target.

3.5. Multiple Reference Graph Reasoning

The subjective biases introduced by people voting for artistic images can have a negative impact on the aesthetic score distribution. Consequently, when predicting the aesthetic quality of artistic images, these biases may influence models, resulting in significant score disparities even for semantically and stylistically similar images. To mitigate the negative effects of subjectivity, we propose the Multiple Reference Graph Reasoning (MRGR) approach, which mimics how people vote during subjective evaluations and leverages related experiences to make more consistent judgments. MRGR incorporates multiple sample initialization and weight-adjusted sampling to learn and establish the inherent relationships among multiple samples, thereby obtaining better artistic aesthetic features.

For multiple reference initialization, we utilize the normalized feature distance values calculated during feature indexing in Sec. 3.4 as the initial weights for multiple samples

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

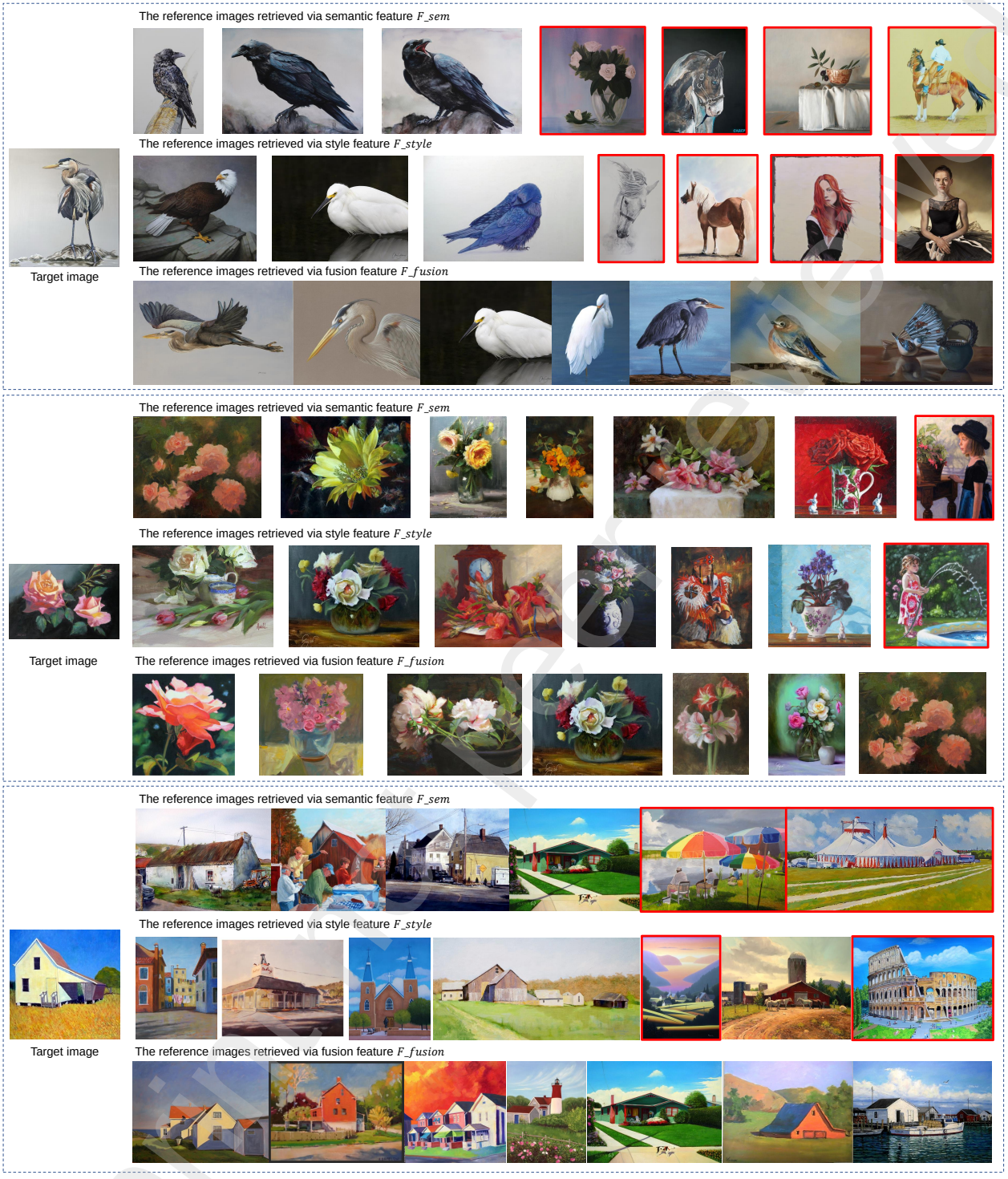
The reference images retrieved via semantic feature F_sem
The reference images retrieved via style feature F_style
Target image
The reference images retrieved via fusion feature F_fusion
The reference images retrieved via semantic feature F_sem
The reference images retrieved via style feature F_style
Target image
The reference images retrieved via fusion feature F_fusion
The reference images retrieved via semantic feature F_sem
The reference images retrieved via style feature F_style
Target image
The reference images retrieved via fusion feature F_fusion

Figure 3: The target image (left) and the reference images (right) formulated by different feature retrieval methods. The first seven images are retrieved for target are shown. The images with red border indicate obvious difference in both semantic content and style compared to the target image.

in the graph network. These weights serve as prior knowledge for defining the initialization relationships among multiple samples. Since smaller feature distances correspond to larger weights, we use their reciprocals as weight values.

With the help of the proposed inter-node initial relationship $\mathbf{A}$ (Eq. 4), we can represent all nodes with both initial edges to learn as a whole graph, *i.e.*, $\{\mathcal{N}, \mathbf{A}, \mathbf{W}\}$, where $\mathcal{N} \in \mathbb{R}^{(M+1)\times d}$ includes target node ($\mathcal{N}_t$) and its M reference nodes ($\mathcal{N}_i$), which have all been represented

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

by d-dimensional feature vectors, and $\mathbf{W}$ is the inter-node relationship to be learned.

$$\mathcal{N}^u \leftarrow \bar{\mathbf{A}}^u \mathcal{N}^{(u-1)} \mathbf{W}^{(u)}, \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^{d \times d}$, $u \in \{1, 2, ..., \mathrm{K}\}$ denotes the current reasoning step number of GCN layer, K is the total iteration number, $\bar{\mathbf{A}}^1 \leftarrow \mathbf{A}$ (Eq. 5), at the end of each reasoning step, $\mathbf{W}^u$ and $\bar{\mathbf{A}}^u$ will be updated respectively by typical error backpropagation.

It is important to note that the MRGR module assumes that multiple samples, such as x_t, x_i, share intrinsic semantic and stylistic similarities, which can more accurately determine through graph reasoning. However, when conducting relationship inference and construction in MRGR, this assumption may only hold for some reference images. Some scenarios and styles may lack corresponding similar reference images in the training set. Hence, when using GCN for inference, the initially weighted relationships are adaptively adjusted by learning artistic attribute features and graph reasoning. We further optimize the training parameters by connecting the adjacency matrices of all samples in the graph structure. The optimized adjacency matrices $\mathbf{A}$ ensure that during the learning process, the weight of target and reference nodes are tailored to learning samples accurately.

Finally, the aesthetic score $\hat{\mathbf{p}}_t$ for the target image can be predicted as follow:

$$\hat{\mathbf{p}}_t = \mathrm{FC}\Big(C\big[\mathcal{R}(\mathcal{N}^{\mathrm{K}})\big]\Big), \tag{6}$$

where $FC(\cdots)$ is a typical fully-connected layer, $C[\cdots]$ is feature concatenation operation, $\mathcal{R}(\cdots)$ is column-wise reshape operation converting its input matrix to a high-dimensional vector (with dimension $(M+1)\times d$). The learning procedure is summarised in Algorithm 1.

3.6. Inference

SSMR computes the reference index for each test sample from the training dataset with the SRIG module. This process typically requires approximately 0.5 hours to complete. Moreover, SSMR requires only one forwarding pass for inference, while multiple samples embedding forward passes are required in our method that utilizes augmentations. The reasoning single-forward inference of the MRGL module is beneficial in subjective-sensitive tasks such as image cropping, image aesthetic assessment and image color enhancement. After reasoning, SSMR determines to predict the aesthetic score via the full connection layer.

4. Experiments

In this section, we first describe the experiment settings used in the reference index and learning pipeline (introduced in Sec. 3.4 and Sec. 3.5), then we elaborate the training and evaluation of the proposed **SSMR** and state-of-the-art AIAA methods on Boldbrush Artistic Image Dataset (BAID), and validate the effectiveness of each module and different settings by exploring the ablation studies.

Algorithm 1 Summary of Key Steps in Our Approach.

Input: target image $\mathbf{I}_t$ and image pool $\mathbf{D}_i$ (Sec. 3.2); pre-trained ImageNet ResNet50 and VGG19 model $\mathcal{H}$; pre-trained Swin-Transformer V2 feature encoder $\mathcal{E}$;

Parameters: reasoning steps N (Sec. 3.5), number of reference image M, value of reference coefficient λ;

1: compute the distance of fusion feature for each image with coefficient λ (Eq. 4);
2: retrieve the similar images of $\mathbf{I}_t$, and select top M among them as the references of $\mathbf{I}_t$ from $\mathbf{I}_i$ (Sec. 3.4);
3: assign inter-node initial relationship $\mathbf{A}$ via Eq. 3;
4: perform N step inter-node reasoning, aiming at obtaining the $\mathbf{W}$ — the weight of inter-node relationship to be learned via Eq. 5 (Sec. 3.5);
5: predict aesthetic score $\hat{\mathbf{p}}_t$ via Eq. 6.

4.1. Experimental Setup

Datasets. We evaluate the proposed approach on BAID [1], VAPS and TAD66K. BAID Yi et al. (2023) consists of 60,337 artistic images covering various art forms, with more than 360,000 votes from online users. It scales the number of votes into the $[0, 10]$ score range, where 0 means the worst and 10 represents the best. Following the common practice in Yi et al. (2023), we split the 60,337 images in BAID into 53,937:6,400 for training and testing, respectively. VAPS [2] Fekete, Pelowski, Specker, Brieber, Rosenberg, and Leder (2022) has 999 fine art paintings from 347 European painters, and we split 200 images for testing and 799 for training. TAD66k [3] He et al. (2022) includes 1,431 artistic images. We keep the same split with the original dataset for selecting 289 images as test and the remaining 1,142 images for training.

Indexing Settings. Our method employs ImageNet pre-trained ResNet50 He, Zhang, Ren, and Sun (2016) to extract 2048-dimensional high-level features as semantic representations and pretrained VGG19 Simonyan and Zisserman (2014) to extract 4096-dimensional high-level features as style representations. We calculate the Euclidean distance between the features of all training sets and the current learning image. To balance the importance of the effect between semantic and style features, we normalize and introduce weight parameters λ. We chose $\lambda = 0.5$ for experimantal comparison, then added two kinds of feature distances. We select the top m samples with the smallest sum of the distances as reference images. Finally, these m reference images exhibit better similarity in terms of semantic and style attributes, which makes them suitable for learning.

Training Settings. We employ the Swin-Transformer V2 Liu, Hu, Lin, Yao, Xie, Wei, Ning, Cao, Zhang, Dong, Wei, and Guo (2022) pre-trained model to extract 1024 dim high-level features from the target and reference images. The feature dimensions for the model input are $((m + 1), 1024)$. We

[1] https://github.com/Dreemurr-T/BAID
[2] https://osf.io/a7xcr/
[3] https://github.com/woshidandan/TANet

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

employ the $smoothL1$ loss function as a constraint to enhance score prediction accuracy. During training, we resize the resolution to 384 and random cropping to 256, adding random image flipping for data augmentation. The training process utilizes the Adam optimizer with a batch size of 32 and a learning rate of 1e-5, decayed by a factor of 0.1 every 2 epochs. To ensure model stability and convergence, we limit the training to 2 GCN layers and 4 reference images representing the optimal configuration. Further details and ablation experiments about the number of GCN layers and reference images can be found in Sec. 4.5 and Sec. 4.6.

Evaluation Metrics. We evaluate AIAA methods on regression and classification tasks. To evaluate the regression performance, we accord to Yi et al. (2023) by adopting two popular evaluation metrics: 1) Spearman's rank correlation coefficient $SRCC$ and 2) Pearson correlation coefficient $PLCC$. We convert the predicted and ground-truth scores to binary-class labels ACC (attractive and unattractive art) using a threshold of 5 (average from 0 to 10) and calculate the accuracy.

4.2. Performance Comparison

Tab. 1 shows the result that compared our method with five SOTA IAA methods and one SOTA AIAA method on the BAID dataset. The IAA methods are NIMA Talebi and Milanfar (2018), MP_{ada} Sheng et al. (2018), MLSP Hosu, Goldlucke, and Saupe (2019), BIAA Zhu, Li, Wu, Zhao, Ding, and Shi (2020), TANet He et al. (2022), and the AIAA method is SAAN Yi et al. (2023). We observe significant performance degradation for the IAA methods NIMA and BIAA on the BAID dataset. For MP_{ada}, MLSP, and TANet, this degradation is particularly due to the difference in theme and style on AVA and BAID datasets. Extracting theme information to evaluate images with different themes by TANet is hard to define in abstract artistic images. Therefore, the IAA methods rely on fine-grained aesthetic score distribution to decrease the robustness of the BAID dataset. Interestingly, SAAN is better in the BAID dataset than IAA methods, consistent with previous reports Yi et al. (2023). The primary reason is that "understanding the style of the artistic image assists in perceiving the aesthetics of the image" and thus it is "especially when there are a wide variety of styles that may have different evaluation criteria."

Our model achieves on **7.4%** and **19.5%** improvement over the best baseline (SAAN) for SRCC and PLCC on the BAID dataset. Compared with IAA methods, our model performs best on all metrics. Compared with the AIAA method that considers the style of the artistic image with a single sample, our framework incorporating the relationship of multiple reference images shows a consistent improvement on all metrics. These results suggest that considering multiple reference images with similar semantics and style can help establish implicit artistic aesthetic perception features for improving the model. Our framework is much more efficient with fewer parameters and outperforms other SOTA methods. The comparision of parameters and running speed is discussed in more detail in Sec. 4.9.

Table 1
Comparison with state-of-the-art open-source IAA and AIAA methods on BAID.

Methods	SRCC↑	PLCC↑	ACC↑
NIMA Talebi and Milanfar (2018)	0.393	0.382	71.01%
MP_{ada} Sheng et al. (2018)	0.437	0.425	74.33%
MLSP Hosu et al. (2019)	0.441	0.430	74.92%
BIAA Zhu et al. (2020)	0.389	0.376	71.61%
TANet He et al. (2022)	0.453	0.437	75.45%
SAAN Yi et al. (2023)	0.473	0.467	76.80%
Ours	**0.508**	**0.558**	**77.72%**

Table 2
Comparison with state-of-the-art open-source IAA and AIAA methods on VAPS and TAD66k datasets.

Dataset	Method	SRCC↑	PLCC↑	ACC↑
VAPS	NIMA Talebi and Milanfar (2018)	0.447	0.455	63.50%
	MP_{ada} Sheng et al. (2018)	0.395	0.404	58.96%
	MLSP Hosu et al. (2019)	0.451	0.464	64.36%
	TANet He et al. (2022)	0.409	0.436	50.50%
	SAAN Yi et al. (2023)	0.469	0.479	**65.82%**
	Ours	**0.482**	**0.501**	64.81%
TAD66k	NIMA Talebi and Milanfar (2018)	0.383	0.408	60.90%
	MP_{ada} Sheng et al. (2018)	0.366	0.382	56.70%
	MLSP Hosu et al. (2019)	0.418	0.422	63.58%
	TANet He et al. (2022)	0.349	0.357	45.32%
	SAAN Yi et al. (2023)	0.425	0.440	65.01%
	Ours	**0.452**	**0.475**	**65.03%**

4.3. Performance on VAPS and TAD66K

Compared to the BAID dataset, both VAPS and TAD66K are smaller in scale, and multiple reference graph learning is expected to enhance the AIAA model in these smaller datasets. We expect that multiple reference learning can demonstrate generality and robustness across various datasets. Therefore, we employ BAID as the image pool to retrieve references for the SSMR model when training the model on the VAPS and TAD66K datasets, utilizing the same backbone. Tab. 2 summarizes the experimental results with the SRCC, PLCC and ACC metrics. For comparison, we retrained five state-of-the-art IAA and AIAA models with public codes, and reported all the metrics under the same experimental setting.

From Tab. 2, it is evident that our SSMR method demonstrates significantly superior performance across all datasets. Our approach achieves a **6.4%** improvement with the same backbone compared to the state-of-the-art model SAAN on TAD66K. These results confirm the generality of multiple reference graph learning. Furthermore, the improved performance on VAPS and TAD66K demonstrates the effectiveness of employing multiple reference graph reasoning on small-scale datasets. These results are also consistent with our expectations.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

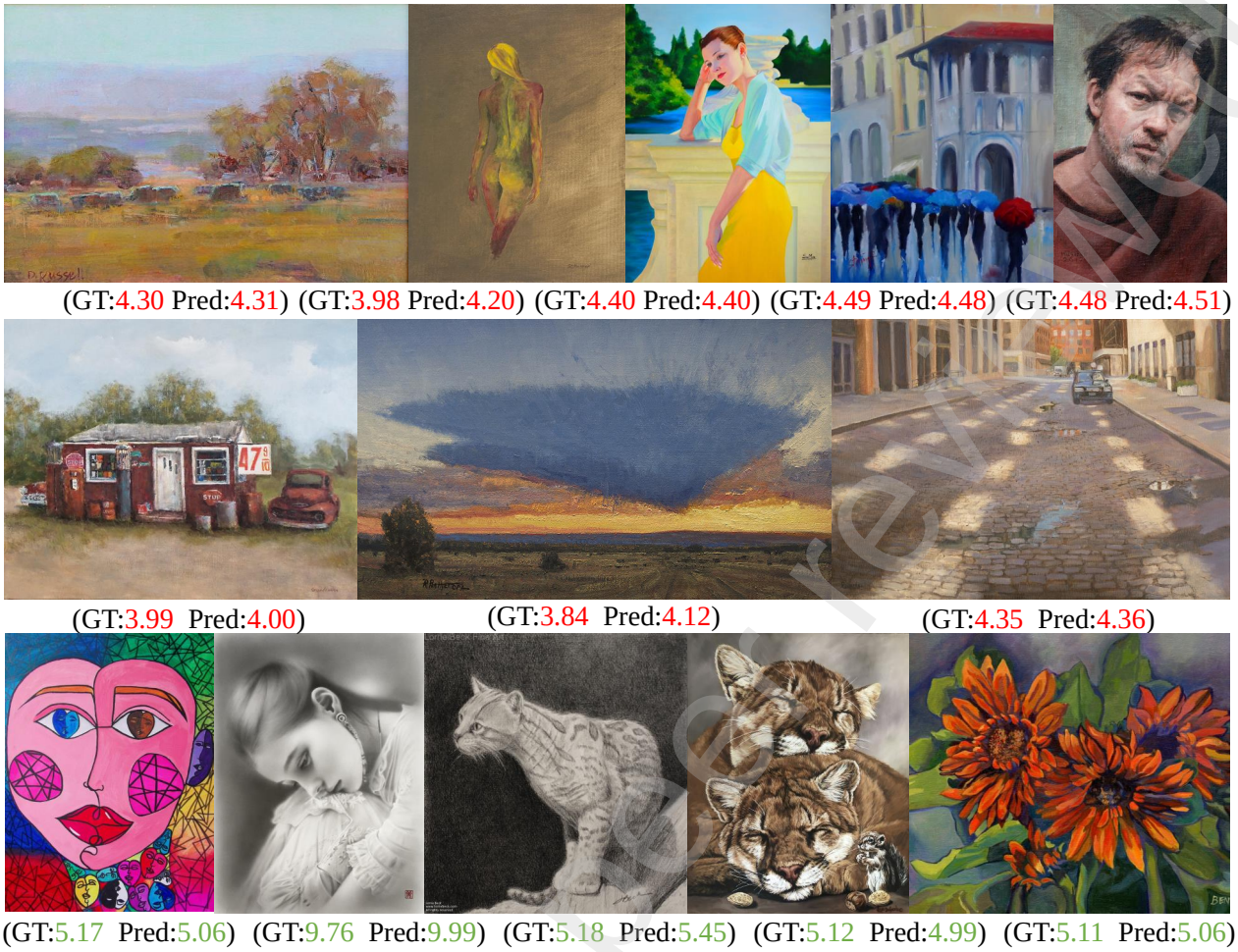

Figure 4: Artistic image aesthetic assessment prediction results. The ground truth (GT) and predicted (Pred) scores are shown underneath each image. The quality is highlighted in different colors, red and green denote low-level and high-level artistic aesthetics, respectively.

4.4. Ablation Study

We explore an ablative study to further investigate the effectiveness of individual components. Tab. 3 shows the result on the BAID dataset. (1) SRIG only: using the SRIG module to retrieve multiple reference images, and then concate all the features and employ the AP module to regress the aesthetic score. Using SRIG only significantly improves evaluation metrics (SRCC from 0.393 to 0.437) over the Baseline, demonstrating the effectiveness of multiple reference images learning. (2) MRGR only: using the MRGR model to learn the relationship of multiple samples, we replace the SRIG module with K-means clustering to select the reference images. We use the MRGR module to reason the reference features of multiple samples show comparable improvement (SRCC from 0.393 to 0.463) with the one sample-only result. These results indicate that multiple reference learning could effectively improve the metrics compared to only single sample learning. (3) SRIG + MRGR: using two modules to generate the index of reference and reasoning the relationship of multiple samples.

These experiments show building semantic and style-based relations in addition to aesthetic features leads to better representation to improve the prediction (SRCC from 0.463 to 0.508) in artistic image aesthetic assessment. The joint use of SRIG and MRGR outperforms using either of them, meaning that our method provides similar reference images with multiple samples for learning the artistic attribution of semantic and style feature to improve the evaluation of all metrics. The joint use of SRIG and MRGR outperforms using either of them, meaning that our method provides similar reference images with multiple images for learning the artistic attribution of semantic and style features to improve the evaluation of all metrics effectively.

4.5. Effect of the number of reference image

We also investigate the effect of the number of reference images for our proposed method. Fig. 5-a shows the result. The smaller the number of reference images is, the **more severe** the artistic aesthetic perception becomes. The error rates of our method deteriorate as m decreases, which shows

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

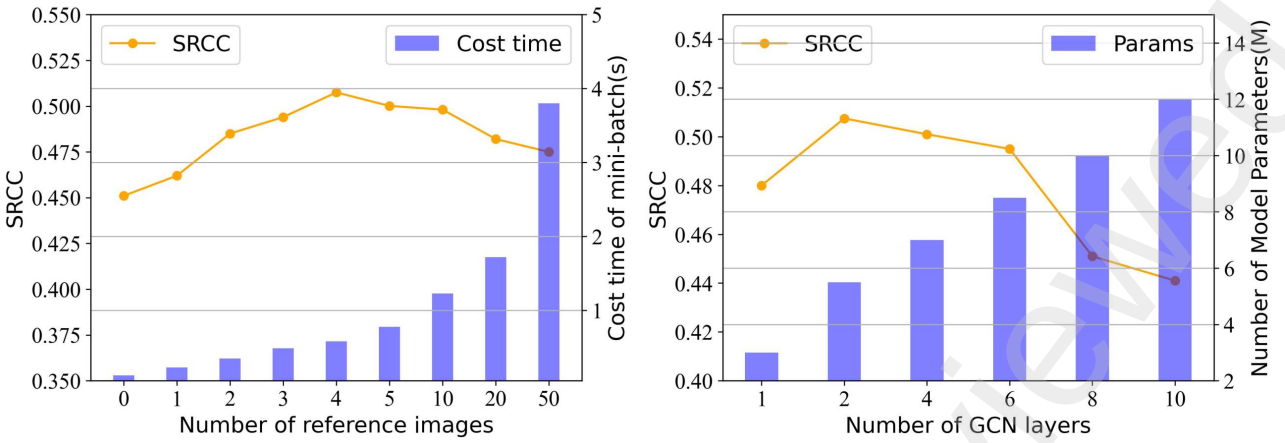

Figure 5: Ablation studies on the number of reference image m (Sec. 4.5) and the number of GCN layer n (Sec. 4.6).

Table 3
Ablation study results on the BAID.

Methods	SRCC↑	PLCC↑	ACC↑
Baseline	0.393	0.382	71.01%
w/o SRIG	0.437	0.425	74.33%
w/o MRGR	0.463	0.471	75.15%
Ours	**0.508**	**0.558**	**77.72%**

that the number of reference images is related to the AIAA method. Fig. 5-a shows the best performance among $m = 4$, indicating its trade-off selection under accuracy and speed for the AIAA task.

4.6. Effect of the number of GCN layer

While we experiment with a widely-used value of the number of GCN layers in SSMR, one might be curious about the impact of the GCN layer on the complexity of the model and the ability of artistic aesthetic representation. Fig. 5-b shows the result with different GCN layers. SSMR is **not much affected** when the GCN layer n is more than 2, while the model's parameters recover performance degradation as the GCN layer increases. These results are because a higher GCN layer can better build relations among multiple reference images. Increasing the GCN layer, however, mitigates multiple sample correlations at the expense of inference latency and adaptation speed. Therefore, we have chosen $n = 2$ as a compromise selection.

4.7. Effect of the coefficient of reference index

In our proposed model, $\lambda = 0.5$ is selected to adjust the coefficient of retrieving reference, which weights the feature distance of similar visual semantics and painting style. Here, we further conduct an experiment to evaluate the performance of SSMR by varying the value of the coefficient, which is achieved by increasing or decreasing the value of the coefficient when retrieving the reference images. The experimental result is shown in Fig. 6.

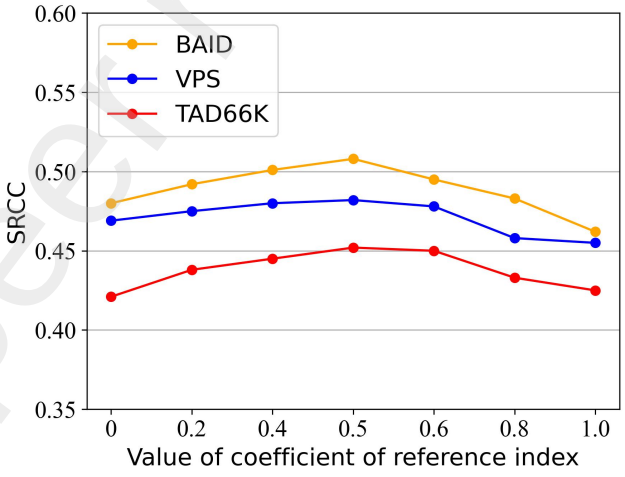

Figure 6: Ablation study on the coefficient of reference index λ (Sec. 4.7).

It is observed from Fig. 6 that the value of coefficient impacts model performance. When the coefficient decreases, the model performance diminishes accordingly, especially for the BAID databset. This results reveals that a smaller value of the coefficient cannot provide sufficient semantic information help model to make aesthetic decision. Conversely, enlarging the value of the coefficient does not achieve improved performance either. These results may be due to the diminishing similarity in considering the feature of artistic style, which leads to increased ambiguity within the artistic style contained in these reference images. In practical implementation, our proposed SSMR model achieves optimal performance when the coefficient is set to 0.5. Consequently, we have adopted the value of coefficient is 0.5 for all our experiments.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

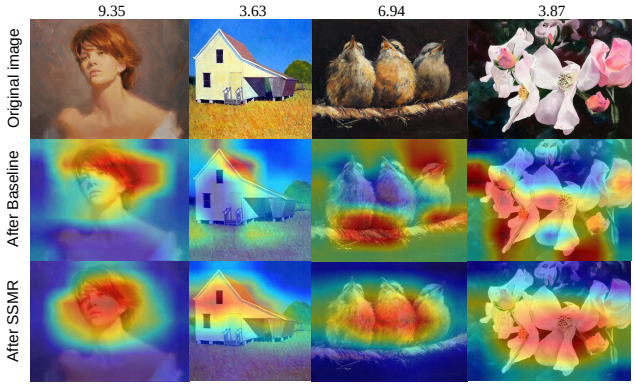

Figure 7: Qualitative results of Class Activation Maps. Given the input image, we extract high-level features and refined features after baseline and after our SSMR model, respectively, and then visualize the feature response via CAM. The ground truth artistic aesthetic score is also given above each input image. Best viewed in color.

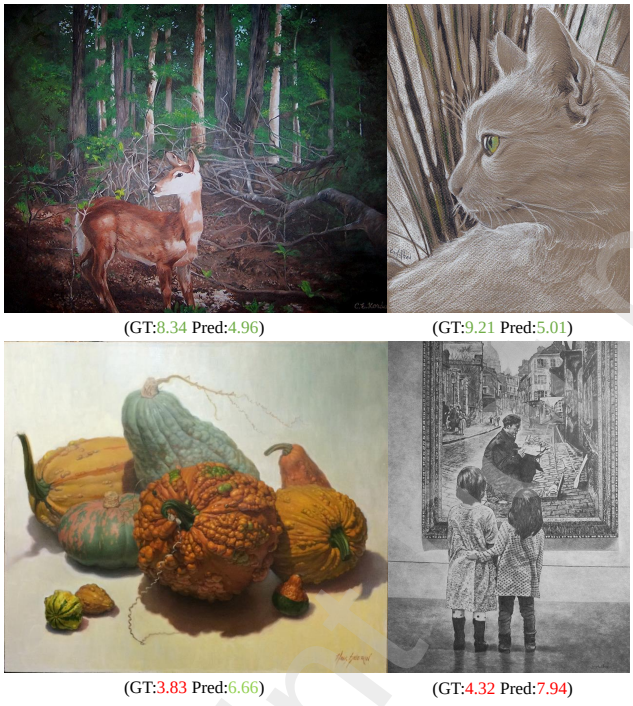

Figure 8: Artistic image aesthetic assessment prediction results. The ground truth (GT) and predicted (Pred) scores are shown under each image. The quality is highlighted in different colors, red and green denote low-level and high-level artistic aesthetics, respectively.

4.8. Visual Analysis

Prediction Results. We show prediction results from the split set of BAID using our proposed SSMR. In Fig. 4, the first two row shows the examples that are predicted with low artistic aesthetic assessment results, and the third row has five examples of high artistic aesthetic assessment results. We find that differences between low-aesthetic and high-aesthetic quality primarily lie in the harmony of the entire image with special semantic (such as beautiful girl and cute animals) and style (realistic and detailed painting style). When the images present attractive semantics, for example, images with concise foreground and details are shown in the third row, the prediction of aesthetics score can be higher. Due to the concise appearance, these images are predicted to be high-aesthetic quality by our model, while the multiple reference images enhances the aesthetic representations mainly due to graph reasoning in the latent space. Besides, we also show several retrieval results by three kinds of feature: F_{sem}, F_{style} and F_{fusion} in Fig. 3, where the reference images retrieved via fusion feature exhibit better semantic and style similarity in the scene of include animals or building.

Table 4
Comparison with state-of-the-art open-source IAA and AIAA on model size and running speed.

Methods	Parameters(M)↓	FLOPs(G)↓	Runtime(ms)↓
NIMA Talebi and Milanfar (2018)	23.51	4.14	**6.2**
MP_{ada} Sheng et al. (2018)	27.56	5.31	9.5
MLSP Hosu et al. (2019)	73.97	32.02	56.7
TANet He et al. (2022)	**13.88**	**2.01**	7.9
SAAN Yi et al. (2023)	30.81	18.25	10.0
Ours	25.62	4.52	6.5

Class Activation Maps. To visualize aesthetics-specific activation within the model, we directly extract Class Activation Maps (CAM) Zhou, Khosla, Lapedriza, Oliva, and Torralba (2016) from our fine-tuned SSMR model. The activation of feature maps after baseline and our SSMR model is shown in Fig. 7. As can be seen, the attended region (highlighted in red) can cover highly correlated objects in the scene after graph convolution, which illustrates that our proposed module refines the semantic features by incorporating multiple reference images inter-information in the network.

4.9. Discussion

Limilation. While our approach accurately predicts the majority of samples, it still encounters failures in some instances. Fig. 8 illustrates instances where the model's predictions significantly deviate from ground truth labels. The figure shows the model's prediction scores alongside corresponding ground truth labels for (a)-(b) close-up animal shots, (c) close-up multiple plant images, and (d) black-and-white portraits. In (a) and (b), the nuanced aesthetic expressions in these anthropomorphic animal close-ups prove challenging to quantify, leading to lower model prediction scores. And (c) and (d) exhibit that close-up images with rich details yield higher model prediction scores. However, (c) lacks aesthetic appeal with rough texture, while (d) suffers from negative expressiveness and clarity due to its monochromatic palette. These results indicate our model's limitation in assessing visual attributes such as texture and color aesthetics.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

Running Speed Comparison. Tab. 4 summarizes the comparative results of various models in terms of parameters and running speed. We conduct the comparison with five code available models. We use images of the same resolution to ensure a fair comparison and employ identical machine configurations. Thanks to our model's simple architecture and low-dimensional input features, it offers a comparable advantage in terms of both FLOPs and running speed. In addition, it's important to note that our approach requires an additional 0.5 hours for feature extraction and reference image indexing before inference. In conclusion, our model's parameter size and running speed are acceptable.

5. Conclusions and Future Work

This paper highlights that artistic image aesthetic assessment often changes across the subjective, and existing IAA and AIAA methods mostly suffer from the single sample learning streams. To address this problem, we present multiple references with a similar semantic and style learning algorithm, SSMR. By performing graph reasoning, the relationships over the artistic aesthetic attribution-based graph are modeled in both the target and reference feature, leading to the fine-grained semantic and style representation. Extensive evaluations show that our proposed SSMR achieves state-of-the-art performance on the benchmark BAID, VAPS and TAD66K artistic image aesthetic dataset. We believe that the insights and findings from this study are a meaningful step toward the practical impact of helping to assess the aesthetics of AIGC's artistic image.

Exploring artistic aesthetic attributes for multiple image reference learning for AIAA is a relatively new and underexplored topic. In this paper, we made some attempts, and there are still some interesting topics worth to explore. First, current datasets typically include only images paired with aesthetic scores, lacking cultural descriptions and commentary labels of artistic images. It is beneficial interpretability for the model to construct an artistic image aesthetic assessment dataset, including cultural background and comment descriptions. Second, evaluating AIAA extends beyond the images and is deeply related to their cultural knowledge. Leveraging knowledge graphs to build cultural knowledge tied to images and employing multimodal techniques that consider both textual and visual data offers the potential for performance improvements. Finally, considering additional artistic aesthetic attributes, such as lighting, color and composition, may enhance artistic aesthetic image representations. These topics would further improve effectiveness or interpretability of artistic image aesthetic models.

References

L. A. Gatys, A. S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2414–2423.

X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1501–1510.

D. Y. Park, K. H. Lee, Arbitrary style transfer with style-attentional networks, in: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5880–5888.

C. Zhang, K. Lei, J. Jia, Y. Ma, Z. Hu, Ai painting: an aesthetic painting generation system, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1231–1233.

M. Gault, An ai-generated artwork won first place at a state fair fine arts competition, and artists are pissed, August 2022.

J.-W. Seow, M.-K. Lim, R. C.-W. Phan, J. K. Liu, A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities, Neurocomputing (2022).

D. Zeng, S. Zhao, J. Zhang, H. Liu, K. Li, Expression-tailored talking face generation with adaptive cross-modal weighting, Neurocomputing 511 (2022) 117–130.

Y. Patotskaya, L. Hoyet, A.-H. Olivier, J. Pettré, K. Zibrek, Avoiding virtual humans in a constrained environment: Exploration of novel behavioural measures, Computers & Graphics 110 (2023) 162–172.

H. Yi, C.-H. P. Huang, S. Tripathi, L. Hering, J. Thies, M. J. Black, Mime: Human-aware 3d scene generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12965–12976.

A. Voynov, K. Aberman, D. Cohen-Or, Sketch-guided text-to-image diffusion models, in: ACM SIGGRAPH 2023 Conference Proceedings, 2023, pp. 1–11.

S.-I. Cheng, Y.-J. Chen, W.-C. Chiu, H.-Y. Tseng, H.-Y. Lee, Adaptively-realistic image generation from stroke and sketch with diffusion model, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4054–4062.

L. Li, T. Zhi, G. Shi, Y. Yang, L. Xu, Y. Li, Y. Guo, Anchor-based knowledge embedding for image aesthetics assessment, Neurocomputing 539 (2023) 126197.

H. Zhang, Y. Luo, L. Zhang, Y. Wu, M. Wang, Z. Shen, Considering three elements of aesthetics: Multi-task self-supervised feature learning for image style classification, Neurocomputing 520 (2023) 262–273.

H. Zeng, Z. Cao, L. Zhang, A. C. Bovik, A unified probabilistic formulation of image aesthetic assessment, IEEE Transactions on Image Processing 29 (2019) 1548–1561.

H. Talebi, P. Milanfar, Nima: Neural image assessment, IEEE transactions on image processing 27 (2018) 3998–4011.

R. Yi, H. Tian, Z. Gu, Y.-K. Lai, P. L. Rosin, Towards artistic image aesthetics assessment: A large-scale dataset and a new method, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22388–22397.

M. Nishiyama, T. Okabe, I. Sato, Y. Sato, Aesthetic quality classification of photographs based on color harmony, in: CVPR 2011, IEEE, 2011, pp. 33–40.

J. Zujovic, L. Gandy, S. Friedman, B. Pardo, T. N. Pappas, Classifying paintings by artistic genre: An analysis of features & classifiers, in: 2009 IEEE International Workshop on Multimedia Signal Processing, IEEE, 2009, pp. 1–5.

J. R. Parker, Algorithms for image processing and computer vision, John Wiley & Sons, 2010.

J. Zhang, Y. Miao, J. Zhang, J. Yu, Inkthetics: a comprehensive computational model for aesthetic evaluation of chinese ink paintings, IEEE Access 8 (2020) 225857–225871.

S. He, Y. Zhang, R. Xie, D. Jiang, A. Ming, Rethinking image aesthetics assessment: Models, datasets and benchmarks, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, 2022, pp. 942–948.

S. Kan, Y. Cen, Y. Li, M. Vladimir, Z. He, Local semantic correlation modeling over graph neural networks for deep feature embedding and image retrieval, IEEE Transactions on Image Processing 31 (2022) 2988–3003.

X. Lu, Z. Lin, H. Jin, J. Yang, J. Z. Wang, Rapid: Rating pictorial aesthetics using deep learning, in: ACM international conference on Multimedia, 2014, pp. 457–466.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919

X. Lu, Z. Lin, X. Shen, R. Mech, J. Z. Wang, Deep multi-patch aggregation network for image style, aesthetics, and quality estimation, in: International Conference on Computer Vision, 2015, pp. 990–998.

S. Ma, J. Liu, C. Wen Chen, A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment, in: Conference on Computer Vision and Pattern Recognition, 2017, pp. 4535–4544.

K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, B.-G. Hu, Attention-based multi-patch aggregation for image aesthetic assessment, in: ACM international conference on Multimedia, 2018, pp. 879–886.

D. Liu, R. Puri, N. Kamath, S. Bhattacharya, Composition-aware image aesthetics assessment, in: Winter Conference on Applications of Computer Vision, 2020, pp. 3569–3578.

D. She, Y.-K. Lai, G. Yi, K. Xu, Hierarchical layout-aware graph convolutional network for unified aesthetics assessment, in: Conference on Computer Vision and Pattern Recognition, 2021, pp. 8475–8484.

S. A. Amirshahi, J. Denzler, Judging aesthetic quality in paintings based on artistic inspired color features, in: 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), IEEE, 2017, pp. 1–8.

X. Guo, T. Kurita, C. M. Asano, A. Asano, Visual complexity assessment of painting images, in: 2013 IEEE International Conference on Image Processing, IEEE, 2013, pp. 388–392.

H. Talebi, E. Amid, P. Milanfar, M. K. Warmuth, Rank-smoothed pairwise learning in perceptual quality assessment, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020, pp. 3413–3417.

D. J. Graham, J. M. Hughes, H. Leder, D. N. Rockmore, Statistics, vision, and the analysis of artistic style, Wiley Interdisciplinary Reviews: Computational Statistics 4 (2012) 115–123.

K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

A. Fekete, M. Pelowski, E. Specker, D. Brieber, R. Rosenberg, H. Leder, The vienna art picture system (vaps): A data set of 999 paintings and subjective ratings for art and aesthetics research., Psychology of Aesthetics, Creativity, and the Arts (2022).

K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12009–12019.

V. Hosu, B. Goldlucke, D. Saupe, Effective aesthetics prediction with multi-level spatially pooled features, in: Conference on Computer Vision and Pattern Recognition, 2019, pp. 9375–9383.

H. Zhu, L. Li, J. Wu, S. Zhao, G. Ding, G. Shi, Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization, IEEE Transactions on Cybernetics (2020).

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4585919